Model adaptation and adaptive training for the recognition of dysarthric speech

Siddharth Sehgal[1], *Stuart Cunningham*[1,2]

[1]Department of Human Communication Sciences, University of Sheffield, Sheffield, United Kingdom
[2]Centre for Asssitive Technology and Connected Healthcare, University of Sheffield, Sheffield, United Kingdom

s.sehgal@sheffield.ac.uk, s.cunningham@sheffield.ac.uk

Abstract

Dysarthria is a neurological speech disorder, which exhibits multi-fold disturbances in the speech production system of an individual and can have a detrimental effect on the speech output. In addition to the data sparseness problems, dysarthric speech is characterised by inconsistencies in the acoustic space making it extremely challenging to model. This paper investigates a variety of baseline speaker independent (SI) systems and its suitability for adaptation. The study also explores the usefulness of speaker adaptive training (SAT) for implicitly annihilating inter-speaker variations in a dysarthric corpus. The paper implements a hybrid MLLR-MAP based approach to adapt the SI and SAT systems. ALL the results reported uses UA-SPEECH dysarthric data. Our best adapted systems gave a significant absolute gain of 11.05% (20.42% relative) over the last published best result in the literature. A statistical analysis performed across various systems and its specific implementation in modelling different dysarthric severity sub-groups, showed that, SAT-adapted systems were more applicable to handle disfluencies of more severe speech and SI systems prepared from typical speech were more apt for modelling speech with low level of severity.

Index Terms: speech recognition, dysarthric speech, speaker adaptation, speaker adaptive training

1. Introduction

Dysarthria is the collective name for a group of motor speech disorders, which result from single or multiple lesions in the brain. It usually results in the loss of motor speech control due to muscular atrophy and incoordination [1, 2]. Across various aetiologies, dysarthric speech is usually characterised by *imprecise consonant production, reduced stress, slow speech rate, hypernasality, harsh and strained voice, muscular rigidity, spasticity, monopitch and limited range of speech movements* [1, 2]. Dysarthria can either be congenital, occurring with conditions such as in cerebral palsy, or acquired, where it develops due conditions such as a stroke or Parkinson's disease.

The effect on speech production of dysarthria is not limited to the musculoskeletal structures, but it can also affect parts of subglottal, laryngeal and supraglottal systems [3]. It usually leads to reduced intelligibility of speech, which can be inversely related to the severity of the underlying condition. On a broad operational scale, severity can be indexed as mild, moderate, severe or any approximation within, such as mild-moderate. For people with severe dysarthria, their speech can be largely unintelligible to unfamiliar listeners.

It is estimated that around 1% of UK population is diagnosed with a neurological disorder each year, although, not all the conditions lead to dysarthria. In UK alone; stroke (416 per 100,000), cerebral palsy (200-300 per 100,000) and Parkinson's disease (200 per 100,000) are amongst the most prevelant causes of motor speech disorders [4, 5].

1.1. Speech interface and dysarthria

Speech has provided an attractive interface for people with dysarthria by enhancing human-human & human-computer interaction. It can enable people with dysarthria to participate in social settings where they can interact with non-familiar communication partners. Moreover, speech as an interface can provide users with a more real-time communication experience to convey messages, in comparison to traditional hardwired switch based interfaces. Earlier studies have shown that systems that deploy automatic speech recognition (ASR) as an interface in a dysarthric setup can have a lower accuracy than hardwired switch-based systems, but, the final message transfer is around 2.5 times faster than the later, even with mis-recognitions followed by corrections [6, 7].

According to a report by [8], more than 70% of dysarthric population with Parkinson's disease or motor neuron disease and around 20% with cerebral palsy or stroke could benefit from some implementation of an augmentative or alternative communication (AAC) device. The benefits of such a setup has proved effective for dysarthric people using speech as an interface for natural communication [9] or enabling them to control physical devices through speech commands [7].

1.2. Automatic speech recognition for dysarthric speech

Dysarthric speech recognition has been investigated for more than two decades [10, 11]. The efficacy of commercial systems has been limited for speakers with mild or mild-moderate dysarthria [12, 13]. In general, decreasing recognition accuracy is linearly related to increasing severity. As a consequence, it has been concluded that the systems are not suited to the higher variability inherent in dysarthric speech.

From a research perspective; acoustic modelling, speaker adaptation and signal enhancement techniques have been explored by researchers to deal with variabilities and disfluencies in dysarthric speech.

The system can be (i) speaker dependent (SD) , which is modelled to recognise only a particular speaker, (ii) speaker independent (SI), which is a generic model map to recognise a range of seen and unseen speakers and (iii) speaker adapted (SA), which attempts to minimise the mismatch between a

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 65–71,
Dresden, Germany, 11 September, 2015.

generic baseline SI model and the intended target speaker. Both generative and discriminative techniques have been exploited to model the acoustics of dysarthric speech. Discriminative approaches like support vector machines has shown some level of success in small vocabulary tasks [14, 15], but by large continuous density HMMs (CDHMM) and its variants remain the most exploited and successful techniques used till date. To get robust model estimates for SD/SI systems, large amounts of training data is usually required. This is not practically viable, since dysarthric speech is afflicted with sparse and inconsistent data problems due to physical constraints, fatigue and muscular atrophy related to a specific individual. Moreover, any dysarthric system will only be effective in real time if the data is collected under conditions where the user will be engaged more often. To overcome this problem to some extent, researchers are using SA systems, which might give SD like performance using lesser amount of data and will be more apt for modelling any unseen user, if a good baseline SI model is available.

Earlier studies using CDHMMs suggested that speaker adapted (SA) systems were suited for mild to moderate dysarthric speakers and speaker dependent (SD) systems better modelled variablities in the severe group of speakers [13, 16]. However, till date there is no common consensus on an established scheme, which indicates the suitability of a technique for a specific type, aetiology or severity of dysarthria. For example, a study by [17], reported a contrary conclusion and suggested that severity is not a good indicator for an optimal selection of modelling approach. Their SA based system outperformed the SD system for most of the speakers used in the study. The disagreement over an optimal approach could also be due to (i) less number of speakers examined in a study, sometimes one, and, (ii) a small vocabulary size, which can create a bias for a certain technique due to the small homogeneous dataset.

1.3. Purpose and aim for the paper

There is a growing need to investigate SA based speech systems, which can be trained with less data and be more accurate for a reasonably large vocabulary. Preparation of SA system usually require using a baseline speaker independent (SI) system and then adapting it using standard techniques. The adaptation methods are usually model based, such as MAP [18] or applies a family of linear transforms, such as MLLR [19]. For dysarthric speech, the basline SI systems are usually prepared from a corpus of typical speech, dysarthric speech or a combination of both.

Although, little work has been done to investigate for an optimal adaptation approach, but some novel attempts have paved the path for further research and investigation. One of the earlier studies comparing SA and SD systems, was reported by [17]. The study was conducted for 7 speakers from the UA-SPEECH database [20] and the results showed that SA system outperformed the SD system for most of the speakers. A more comprehensive study was conducted by [21] on the same dataset that included all the speakers in the UA-SPEECH corpus. They tested a SD system alongside a MAP based SA system. An array of SI baseline models were used for adaptation purposes. Firstly the study showed an average relative increase of 34.5% over the earlier reported results by [17]. Secondly, the results showed that SI system using all the dysarthric speech data forms the best baseline system for MAP adaptation. To the best of our knowledge, the results reported by [21] seems to be the best till date on a relatively large vocabulary size of 255 words for a particular dysarthria type covering a range of severities.

This paper builds up upon these earlier studies and (i) investigates the best SI baseline system for adaptation of dysarthric speech, (ii) explores hybrid adaptation approach using MLLR-MAP and (iii) investigate the efficacy of speaker adaptive training (SAT) [22] to implicitly annihilate the inter-speaker variabilities during the training process.

In the paper, section 2 will detail about the data preparation and methodology used for the experiments, section 3 will present and analyse the recognition results, section 4 will put some collective discussion for the results and section 5 will have the concluding remarks and considerations for the future work.

2. Experimental Setup

2.1. Data preparation

All the experiments presented in this paper used two standard corpora for typical speech, viz., WSJ0 SI-84 [23] that consists of read speech from 84 North American english speakers with texts drawn from a machine-readable corpus of Wall Street Journal news, and, WSJCAM0 [24] , which is a British english version of WSJ database that consists of data from 92 training speakers. For WSJCAM0, data was also included for speakers from the development and two evaluation test sets.

In addition, UA-SPEECH [20] corpus was used, which consists of data from 15 dysarthric speakers with cerebral palsy and 13 control speakers. There are 765 isolated words (455 distinct) per speaker collected in three separate blocks, where each block consists of 10 digits, 26 international radio alphabets, 19 computer commands, 100 common words and 100 distinct uncommon words, which were not repeated across blocks. In addition, the corpus also provides a rough estimate of perceptual speech intelligibility ratings for each dysarthric speaker by five naive listeners. The ratings given will be used in all the experiments for ordering the speakers in various severity groups. All the

Corpus	Speakers	Training Files
WSJ SI-84	84	14377
WSJCAM0 †	136	18537
UA-CTL	13	41819
UA-DYS	15	44277

Table 1: *A summary of each training corpus in the system. UA-CTL and UA-DYS codes are used for UA-SPEECH control and dysarthric speakers. (†) Four evaluation speakers with no secondary microphone data were excluded from WSJCAM0.*

block one (B1) and block three (B3) data from UA-SPEECH was used for training & adaptation purposes and block two (B2) was solely used for all the reported test results in the paper. Because dysarthric speakers can take a longer duration to utter words, the UA-SPEECH training data had to be logically resegmented to get rid of extra silences around word boundaries. Only 200 ms of silence was appended to either side of the word for training. However, test data block B2 was left untouched to maintain the natural speaking conditions. Data from all the microphones was used for each corpus for training and adaptation purpose and a summary is given in Table 1.

For acoustic modelling, data from all the corpora was processed as 12 dimensional MFCC features with c_0 and cepstral mean normalisation. First and second order time derivatives were also appended giving a 39 dimensional feature vector per frame. Speech was analysed in 25 ms window with a 10 ms target shift rate.

2.2. Acoustic Modelling

The continuous density HMM in all the experiments are word-internal tied-state triphone models with clustering performed using phonetic decision trees. It follows a strict left-to-right topology with 16 Gaussian components used per state. Silence states were modelled using 32 Gaussian components.

2.3. Methodology

One of the aim of the paper is to test the efficacy of a good baseline SI system that is more apt for adaptation purposes. This is an extension of the SI systems that was described in [21]. Table 2 summarises the SI systems that were constructed for adaptation purposes.

System Code	Training Dataset Used
SI-00	WSJ SI-84 + WSJCAM0
SI-01	UA-DYS excluding target test speaker
SI-02	UA-DYS
SI-03	UA-CTL
SAT	UA-DYS

Table 2: *Summary of baseline systems and the corpus used for its preparation.*

The SI systems intrinsically model the speaker characteristics and acoustic realisations in speech, which are considered constant throught the database. During typical speaker adaptation, the optimal model set $\tilde{\Phi}$, given a set of S speakers in the system is generally represented as:

$$\tilde{\Phi} = \arg\max_{\phi} \mathcal{L}(O;\phi) = \arg\max_{\phi} \prod_{s=1}^{S} \mathcal{L}(O^{(s)};\phi)$$

where $\mathcal{L}(O^{(s)};\phi)$ is the likelihood of the observation sequences from speaker s, given the current set of model estimates ϕ.

In addition to various SI systems, SAT modelling was also considered in the current study, which splits information into various homogeneous blocks, e.g. data pertaining to a particular speaker for incorporating speaker induced variations. SAT training uses two sets of parameters, a canonical model ϕ_c, usually hypothesised to represent phonetically relevant speech variabilities, and the set of transforms $\mathcal{T}^{(s)}$ to represent the speaker variabilties. This is given as:

$$(\tilde{\Phi_c}, \tilde{\mathcal{T}}) = \arg\max_{(\phi_c,\mathcal{T})} \prod_{s=1}^{S} \mathcal{L}(O^{(s)};\mathcal{T}^{(s)}(\phi_c))$$

In the above equation speaker induced variations are modelled by $\mathcal{T}$ and the canonical model is updated, given each transform. The entire SAT paradigm works iteratively in an interleaved fashion and can be depicted as shown in figure 1.

SAT based on MLLR transforms should be able generate robust canonical model estimates, however, it comes with computational and memory overheads [25], making it impractical for implementation. Such issues are usually avoided by applying constrained MLLR (CMLLR) [26, 27], which uses the same transform for both means and variances. The transforms are computed for each homogeneous block of data. SAT with CMLLR results in a kind of feature normalisation during model training and have the same computational load as any other standard HMM update. Unlike SI models which can be directly

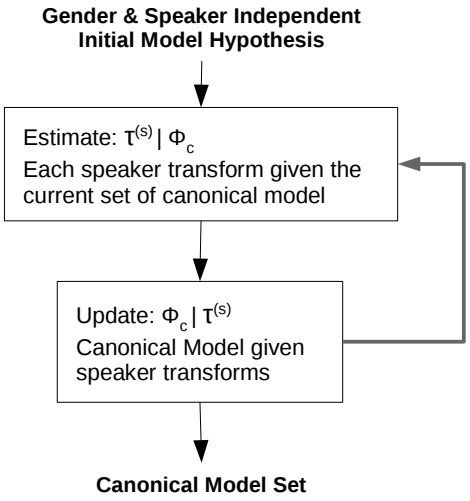

Figure 1: An overview of the SAT framework

used for recognition, SAT canonical model sets are not suited for direct decoding. Both systems are usually adapted to some target test condition.

In this paper, we present the results of the SI and SAT models using MLLR, MAP and MLLR-MAP based adaptation techniques. SAT canonical models are intentionally trained using only UA-DYS speakers to implicitly reduce the inter-speaker variabilities associated with dysarthric speech in general across varying degree of severities. The MLLR implemented uses a two-pass static adaptation procedure. The first pass performs a global transformation and the second pass uses the global transforms to produce more accurate transforms using a regression class tree with 32 terminal leaf nodes.

3. Results

All the test results presented in the paper are obtained on test set B2 of the UA-SPEECH corpus. Since the database comprises of single word utterances, the decoding grammar was strictly restricted to recognise only one of the possible test words, mostly preceded and succeeded by silences. There are 255 distinct competing words in the test block with a total of 22281 files from all speakers and microphones.

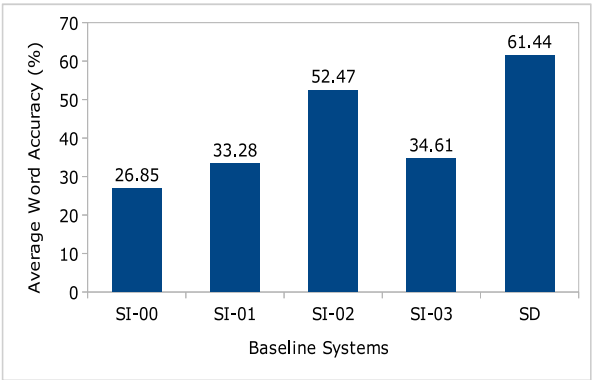

Figure 2: *Average word accuracy for the baseline SI systems along with the SD result.*

3.1. Baseline Systems

The first set of experiments involved obtaining recognition scores of all the baseline SI systems. These were then compared with the SD performance. Figure 2 shows the average baseline accuracy of all the SI systems. SI-00 has the lowest baseline result, which can be explained by the fact it was training only on typical speech. The high accuracy was obtained using the SI-02 system, which was trained on the largest amount of dysarthric speech data.

3.2. Baseline Adapted Systems

All of the baseline systems were adapted for each test speaker. Standard techniques were used and the results are shown in Figure 3. MAP clearly outperforms the MLLR based adaptation for all the systems except SI-00. This may be an example of non-informative priors. The SI-00 models are trained from WSJ0 + WSJCAM0 datasets, which contains only typical speech, and therefore presents no useful information about the model parameter distributions of the adaptation and test datasets.

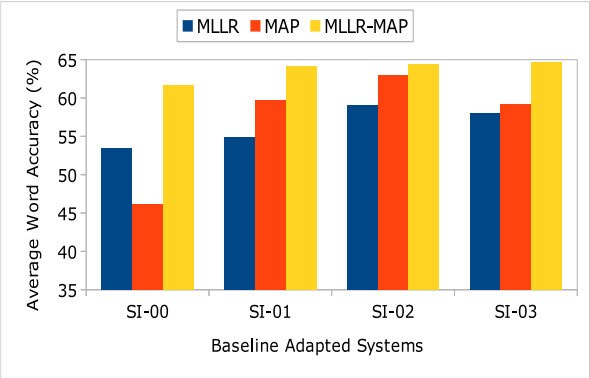

Figure 3: *Adaptation scores for the baseline SI systems.*

Following on from this observation we implemented a combined approach that involves generating MLLR transforms for the target speaker followed by MAP adaptation. By doing this, MLLR adapted parameters can act as informative priors for the MAP process. For all the SI systems, the MLLR-MAP combination outperformed all other adaptation approaches. For this reason the remainder of the paper will primarily focus on results obtained using a MLLR-MAP approach.

Intuitively, it may be thought that SI-01 or SI-02 should form an optimal set of baseline models for adaptation, since they exhibit less difference between the training, adapted and test conditions. Overall, the best MLLR-MAP scores for dysarthria and typical speech based SI systems was found to be for SI-02 and SI-03.

3.3. SAT-adapted vs Other Systems

One of the aims of the paper is to study the effect of SAT based modelling to reduce inter-speaker variations during training time. This section reports SAT-adapted results and compares it to the state-of-the-art SD system and other SI-adapted systems reported earlier. Figure 4 gives a comparison of the MLLR-MAP based SI and SAT systems. Clearly, SAT-adapted model sets outperform all the other tested systems

It should be noted that SD system performs poorer than all the other adapted systems. Indeed, it can be seen in Table 3 that SD system does not perform better than any of the

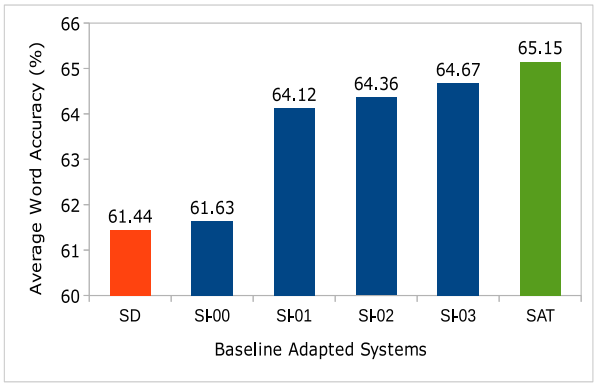

Figure 4: *Comparison of SD and MLLR-MAP based SI & SAT systems.*

SA systems (*except one speaker*) under various intelligibility sub-groups. This gives us an average understanding that adaptation can be an effective approach to model dysarthric speech of varying severities. A similar finding about the efficacy of SA systems was also reported in a study by [17]. Our findings are contrary to some of the earlier published results [16, 13], which were more inclined to favour SD systems with increasing severity. In another study by [21], SI systems prepared from only dysarthric datasets produced better adapted models for most of the speakers.

In contrast our findings suggest that SI systems like SI-03, prepared from typical speech can also adapt as well as a dysarthric speech-based SI system. In order to justify our presumption, the effectiveness of all the MLLR-MAP based SAT and SI systems along with SD system was statistically analysed using Cochran's Q test. All the systems were tested for differences across all the test speakers. The null hypothesis was rejected at $\alpha = 0.01$, *degrees of freedom* = *5*, which meant that all the systems were not equally effective for modelling dysartric speech in general. Later a pairwise Cochran's Q test was conducted between the system with the best absolute average score (SAT) and all others. The test showed that SAT was significantly different to all other systems at $p < 0.01$, except for the SI-03 system.

3.4. Severity Based System Results

So far we have reported all our findings averaged across all the test speakers regardless of the severity. However, to have a more customised approach for preparing systems for specific speakers it is important to individually study the effect of SD and SA based systems under various severity groups. The MLLR-MAP results reported earlier were investigated further for each of the different severity groups. Figure 5 gives an overall picture of how the baseline SI systems performed for various intelligibility sub-groups and Figure 6 shows the effect of adaptating the respective baseline systems along with SAT estimates. The speakers at the lowest intelligibility group showed inclination towards SAT based system or systems prepared with some dysarthric data, while, speakers in the highest intelligibility group benefitted from the presence of only typical speech data. Table 3 gives a detailed test report for all the UA-DYS speakers.

In order to understand differences between the systems, a Cochran's Q test was again applied to study the system differences under various speaker severity groups. The summary of the results of this test are shown in Table 4. It shows that SAT

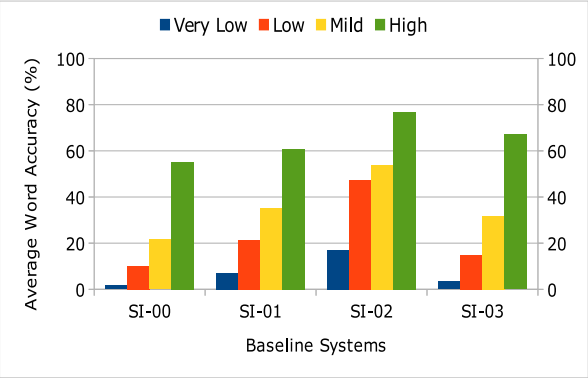

Figure 5: *Word accuracy for the baseline SI systems under various intelligibility groups (Very Low, Low, Mild, High).*

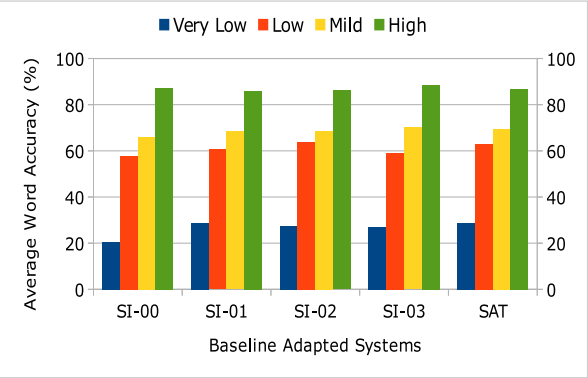

Figure 6: *MLLR-MAP scores for the SAT & SI systems under various intelligibility groups (Very Low, Low, Mild, High).*

Intelligibility	Speaker	SD	MLLR-MAP				
			SI-00	SI-01	SI-02	SI-03	SAT
Very Low	M04 (2%)	6.54	8.98	9.5	8.54	8.11	9.68
	F03 (6%)	32	27.61	37.49	36.01	36.81	38.36
	M12 (7%)	32.24	17.76	35.08	32.31	30.71	32.9
	M01 (17%)	16.76	27.03	28.32	28.22	27.46	29.22
Sub Acc.		23.52	20.61	28.82	27.36	26.95	28.71
Low	M07 (28%)	62.33	69.7	69.26	68.89	61.91	66.06
	F02 (29%)	61.08	37.62	50.12	54.02	50.93	56.93
	M16 (43%)	64.29	68.08	62.76	66.47	65.23	66.55
Sub Acc.		62.48	57.89	60.56	62.92	59.03	62.98
Mild	M05 (58%)	70.48	64.27	69.93	70.6	67.47	71.83
	M11 (62%)	58.18	56.57	63.8	66.06	68.1	65.62
	F04 (62%)	62.66	76.06	70.57	68.48	74.52	70.57
Sub Acc.		64.44	66.12	68.34	68.51	70.13	69.54
High	M09 (86%)	80.96	83.11	84.43	85.62	87.82	86
	M14 (90%)	77.76	80.4	80.09	79.2	85.71	80.84
	M10 (93%)	84.28	91.77	86.28	87.21	91.33	88.08
	M08 (95%)	85.86	87.96	87.21	86.47	87.4	87.34
	F05 (95%)	86.46	92.14	92.01	92.33	90.58	92.08
Sub Acc.		83.07	87.08	86.01	86.17	88.57	86.87
Overall Acc.		61.44	61.63	64.12	64.36	64.67	65.15

Table 3: *Average word accuracy rates for SD and all SI baseline systems adapted using MLLR-MAP. The table also shows sub accuracy scores under various intelligibility groups. The best scores are highlighted in grey for each row.*

system is statistically equivalent to some other systems in the *very-low, low and mild* sub-group of speakers.

Intelligibility	Best performing systems ($p < 0.05$)
Very Low	SAT, SI-01
Low	SAT, SD, SI-02
Mild	SAT, SI-03
High	SI-03

Table 4: *Cochran's Q analysis for all the systems under various intelligibility sub-groups.*

For the *high* intelligibility sub-group, system trained from typical speech data with similar recording and vocabulary setup as the test dysarthric conditions was significantly different to all the other competing systems.

4. Discussions

The results reported in Section 3 show that it is difficult to train a system to model the variabilities in dysarthric speech and to generalise to speakers of different severities. For example, when studying the performance of various baseline systems in section 3.1, it was interesting to note that SI-03 had similar performance to SI-01 system, despite being trained from typical speech data. We think that SI-03 models will be making use of information from homogeneous vocabulary and recording conditions as the test dysarthric conditions.

The findings also show that SD system were not the most effective to model dysarthric speech. This can be partially attributed to the relatively small amount of data per speaker in UA-SPEECH, especially when compared to previous studies in the literature [16, 13]. The test block B2 also comes with many unseen acoustic realisations in the form of 100 unique "uncommon words" and an SD system is usually only tuned to maximise the model fit for the seen data blocks during training. In contrast, a SA system might overcome this problem to some

extent by using acoustic information present from other users in the baseline SI systems. This might be a contributing factor for all the adapted systems to be significantly better than SD system.

Another point of interest, reported in section 3.3, indicated that to model dysarthric speech in general, SAT and SI-03 systems were not significantly different. Hence the selection of a good baseline system to adapt from cannot depend on any particular dataset. It needs a more thorough investigation to understand the acoustics of dysarthric speech at an intra and inter speaker level. For instance, these results suggest that the variabilities in dysarthric speech can be better accommodated from modelling both typical and dysarthric domains. One such attempt was reported by [28], where background interpolation MAP was implemented to obtain an intermediate prior acoustic model to narrow the gap between two disparate SI systems (*typical & dysarthric*), albeit, the reported results were no better than those reported by [21]. Our best overall results, as reported in sections 3.3 & 3.4, are based on MLLR-MAP adapted SAT systems. It gives an absolute gain of 22.91% (54.36% relative) over results of [28] and an absolute gain of 11.05% (20.42% relative) over results of [21].

The choice of a particular system for a given target speaker is not completely clear, even when analysis is carried out at specific intelligibility levels. Table 4 indicates several possible choices in the lower intelligibility group of speakers. Since dysarthric speech will be more variable in the lower intelligibility group, the presence of SI-01 and SI-02 does not come in as a surprise as they will be inherently capable of modelling some of the common disfluencies. Although, the presence of SD system in the *low* intelligibility sub-group might suggest some corpus bias towards a particular speaker. It would appear that the choice of a baseline model for a particular target speaker may be determined by the amount of training data available.

Despite the fact that several alternatives appear to be equivalent for different groups of speakers, it is noticeable that SAT-based systems are among the best performing for the very low to mild groups of speakers. This may be due to the implicit capability of SAT to remove the speaker induced variations during training time. This speaker normalising might be having a nullifying effect on some complex variabilities present across all the speakers.

Among systems trained with typical speech, SI-03 is significantly a better base model for adaptation than SI-00. This is despite being trained with a smaller dataset. This may suggest that large quantities of typical speech data might not be necessary for the base models adapted to recognise dysarthric speech.

Lastly, as shown in Table 4, it is not surprising to observe that SI-03 was the best performing system for speakers with a high intelligibility. Perceptually, high intelligibility dysarthric speech is more akin to typical speech. Table 3 clearly shows the inclination of typical speech baseline systems (*SI-00, SI-03*) to model *high* intelligibility sub-group of speakers. In addition to acoustic similarities, as mentioned earlier, SI-03 system also has an additional benefit of homogeneous vocabulary and recording conditions.

5. Conclusions and future work

The current paper investigated the effectiveness of SAT-adapted, SD and SI-adapted systems to model dysarthric speech. We found that the hybrid MLLR-MAP based technique outperformed other adaptation procedures. All the MLLR-MAP based SAT and SI systems produced an absolute gain over similar results reported in earlier studies [21, 28] for this corpus. SAT-adapted systems had the highest overall average word accuracy for all dysarthric speakers. Although, systems trained from typical speech data with homogeneous recording conditions and vocabularies as the test dysarthric conditions were not significantly different to SAT-adapted systems.

It is difficult to assert at this time about the best strategy of SI or SAT based systems for robust adaptation and recognition of a target dysarthric speaker. SAT-adapted systems can implicitly model inter-speaker variabilities and proved to be significantly better at recognising speech from speakers with lower intelligibility. in contrast, typical speech systems were more inclined to model high intelligibility sub-group of speakers. The results also showed that that adaptation might be a better than corresponding SD systems to model dysarthric speech.

Despite the results reported here, there is still no consensus on the best approach to model dysarthric speech with varying severity, aetiology or type. Future work should investigate the SAT-based modelling approach, especially approaches for customising baseline systems prior to adaptation to a specific speaker.

6. Acknowledgements

This report is independent research funded by the National Institute for Health Research Invention for Innovation programme, Speech Therapy Apps for Rehabilitation, (II-LB-0214-20008). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

7. References

[1] F. Darley, A. Aronson, and J. Brown, "Clusters of deviant speech dimensions in the dysarthrias," *Journal of Speech and Hearing Research*, vol. 12, pp. 462–496, 1969.

[2] J. Duffy, *Motor Speech Disorders : Substrates, Differential Diagnosis, and Management*, 2nd ed. Elsevier Mosby, 2005.

[3] R. Kent, J. Kent, G. Weismer, and J. Duffy, "What dysarthria can tell us about the neural control of speech," *Journal of Phonetics*, vol. 28, no. 3, pp. 273–302, 2000.

[4] RCSLT, *Communicating Quality 3: RCSLT's Guidance on Best Practice in Service Organisation and Provision*. Royal College of Speech & Language Therapists, 2006. [Online]. Available: http://books.google.co.uk/books?id=udcuAAAACAAJ

[5] "Resource manual for commissioning and planning services for slcn," http://www.rcslt.org/speech_and_language_therapy/commissioning/aac_plus_intro, 2009, online; accessed on: 13-May-2015.

[6] M. S. Hawley, "Speech recognition as an input to electronic assistive technology," *The British Journal Of Occupational Therapy*, vol. 65, no. 1, pp. 15–20, 2002.

[7] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer, "A speech-controlled environmental control system for people with severe dysarthria." *Med Eng Phys*, vol. 29, no. 5, pp. 586–593, 2007.

[8] "Communication matters research matters: an aac evidence base," http://www.communicationmatters.org.uk/beyond-the-anecdote, 2013, online; accessed on: 13-May-2015.

[9] M. Hawley, S. Cunningham, P. Green, P. Enderby, R. Palmer, S. Sehgal, and P. O'Neill, "A voice-input voice-output communication aid for people with severe speech impairment," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 21, no. 1, pp. 23–31, 2013.

[10] M. Fried-Oken, "Voice recognition device as a computer interface for motor and speech impaired people," *Archives of Physical Medicine and Rehabilitation*, vol. 66, no. 10, pp. 678–681, 1985.

[11] C. Coleman and L. Meyers, "Computer recognition of the speech of adults with cerebral palsy and dysarthria," *Augmentative and Alternative Communication*, vol. 7, no. 1, pp. 34–42, 1991.

[12] K. Hux, J. Erickson, N. Manasse, and E. Lauritzen, "Accuracy of three speech recognition systems: Case study of dysarthric speech," *Augmentative and Alternative Communication*, vol. 16, pp. 186–196, 2000.

[13] P. Raghavendra, E. Rosengren, and S. Hunnicutt, "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems," *AAC: Augmentative and Alternative Communication*, vol. 17, no. 4, pp. 265–275, 2001.

[14] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009.*, 2009, pp. 4605–4608.

[15] V. Wan and J. Carmichael, "Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data," in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, 2005, pp. 3321–3324.

[16] F. Rudzicz, "Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech," in *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, ser. Assets '07, 2007, pp. 255–256.

[17] H. Sharma and M. Hasegawa-Johnson, "State-transition interpolation and map adaptation for hmm-based dysarthric speech recognition," in *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, 2010, pp. 72–79.

[18] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[19] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[20] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, 2008, pp. 1741–1744.

[21] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, vol. 2, 2012, pp. 1774–1777.

[22] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Fourth International Conference on Spoken Language, ICSLP 96., Proceedings.*, vol. 2, 1996, pp. 1137–1140.

[23] D. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the Workshop on Speech and Natural Language*, ser. HLT '91, 1992, pp. 357–362.

[24] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-95.,*, vol. 1, 1995, pp. 81–84.

[25] M. Spyros, S. Rich, J. Hubert, and N. Long, "Practical implementations of speaker-adaptive training," in *DARPA Speech Recognition Workshop*, 1997.

[26] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, 1995.

[27] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[28] H. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech and Language*, vol. 27, no. 6, pp. 1147–1162, 2013.